Date: 10/20/00 Express Mail Label No. EL552572736US

Inventors: Thomas J. Hudson, James Engert and Andrea Richter

Attorney's Docket No.: 2825.1021-003

IDENTIFICATION OF ARSACS MUTATIONS AND METHODS OF USE THEREFOR

RELATED APPLICATIONS

This application claims the benefit of U.S. Provisional application Serial No. 60/160,588, filed October 20, 1999, the entire teachings of which are incorporated herein by reference.

BACKGROUND OF THE INVENTION

Autosomal recessive spastic ataxia of Charlevoix-Saguenay (ARSACS) is an early-onset neurodegenerative disease with high prevalence in the Charlevoix-Saguenay-Lac-Saint-Jean (CSLSJ) region of Quebec. Disease progression is rapid through young adulthood, with most patients requiring wheelchairs by their early forties. The disease is characterized by abolished sensory nerve conduction, reduced motor nerve velocity, and a unique clinical feature of hypermyelination of retinal nerve fibers. Additional pathological features include atrophy of the upper cerebellar vermis, absence of Purkinje cells, and possibly abnormal neuronal lipid storage (Bouchard, J-P., *In*: Handbook of Clinical Neurology 16: Hereditary neuropathies and spinocerebellar degenerations, J.M.B.V. de Jong, Ed., pp. 451-459, Elsevier Science Publishers, Amsterdam (1991)). A developmental defect in the myelination of both retinal and peripheral nerve fibers has been proposed as the physiological basis of the disease (Bouchard, J-P., *et al., Neuromuscular Disorders* 8:474-479 (1998)). More than 300

patients have been identified, and the estimated carrier frequency is 1 in 22 in the Charlevoix-Saguenay-Lac-Saint-Jean (CSLSJ) population of northeastern Quebec (3).

SUMMARY OF THE INVENTION

As described herein, the ARSACS gene, referred to herein as "*spastin*" (also known as *sacsin*), has been mapped to chromosome 13q11 by linkage analysis and cloned from human, mouse and hamster. The gene was identified by using fine-structure linkage disequilibrium (LD) mapping to narrow the disease interval and then performing sample-sequencing to identify candidate genes. The *spastin* gene has a remarkable feature in that it contains a large exon spanning at least 12,794 base pairs of genomic DNA and comprises an open-reading frame of 11,487 base pairs. As described herein the gene is highly conserved in mouse. This exon of *spastin* is the largest found in any vertebrate organism. The deduced protein contains three large domains with sequence similarity to each other, as well as to the protein predicted to be encoded by an open reading frame identified in *Arabidopsis* genomic DNA. These domains contain a subdomain with sequence similarity to heat-shock proteins, suggesting a role in chaperone-mediated protein folding. *Spastin* appears to be expressed in a wide variety of tissues including brain and central nervous system. Alterations in the *spastin* gene have been identified as described herein which correlate strongly with ARSACS, including at least two alterations which have severe effects on the encoded protein, providing strong evidence that mutations in the open reading frame of the *spastin* gene are responsible for ARSACS.

The present invention relates to an isolated nucleic acid molecule comprising a *spastin* gene or portion of said gene as described herein. In one embodiment, the invention relates to an isolated nucleic acid molecule comprising a nucleotide sequence selected from the group consisting of SEQ ID NOS: 1, 3, 7, 9, 11, 12, 13, 14 and 15 and the complement of SEQ ID NOS: 1, 3, 7, 9, 11, 12, 13, 14 and 15. In another embodiment the invention relates to an isolated nucleic acid molecule comprising an exon from a vertebrate gene wherein said exon is at least 1150 base pairs in length. The

invention also relates to an isolated nucleic acid molecule consisting of a nucleotide sequence selected from the group consisting of SEQ ID NOS: 1, 3, 7, 9, 11, 12, 13, 14 and 15 and the complement of SEQ ID NOS: 1, 3, 7, 9, 11, 12, 13, 14 and 15. In a preferred embodiment the genes of the invention are human genes. The invention also relates to an isolated nucleic acid molecule consisting of a nucleotide sequence selected from the group consisting of SEQ ID NOS: 21-66 and the complement of SEQ ID NOS: 21-66.

The present invention also includes fragments of the *spastin* genes described herein. For example, the invention relates to an isolated portion of a nucleic acid sequence selected from the group consisting of SEQ ID NOS: 1, 3, 7, 9, 11, 12, 13, 14 and 15 and the complement of SEQ ID NOS: 1, 3, 7, 9, 11, 12, 13, 14 and 15, wherein the portion is at least about 10 nucleotides in length.

The invention also relates to nucleic acid molecules having substantial sequence identity to the specific sequences disclosed herein. In one embodiment, the invention relates to a nucleic acid molecule comprising a nucleotide sequence which is at least about 60% identical to a nucleotide sequence selected from the group consisting of SEQ ID NOS: 1, 3, 7, 9, 11, 12, 13, 14 and 15 and the complement of SEQ ID NOS: 1, 3, 7, 9, 11, 12, 13, 14 and 15. In another embodiment, the invention relates to a nucleic acid molecule which hybridizes under high stringency conditions to a nucleotide sequence selected from the group consisting of SEQ ID NOS: 1, 3, 7, 9, 11, 12, 13, 14 and 15 and the complement of SEQ ID NOS: 1, 3, 7, 9, 11, 12, 13, 14 and 15.

The nucleic acid molecules of the present invention, or portions thereof, can be used as probes to isolate and/or clone substantially similar or functionally equivalent homologues of the *spastin* family of genes. The polynucleotides of the present invention can also be used as probes to detect and or measure expression of the genes encoded by the present invention. The probes of the present invention can be DNA, RNA or PNA. Expression assays, such as Southern blot analysis and whole mount *in situ* hybridization, are well known in the art. The polynucleotides of the present

invention, or portions thereof, can also be used as primers to clone homologues or family members by PCR using techniques well known in the art.

The invention further relates to nucleic acid constructs comprising the isolated nucleic acid molecules of the invention, as well as to a recombinant host cell comprising the isolated nucleic acid molecules of the invention. The invention further relates to a method for preparing a polypeptide encoded by an isolated nucleic acid molecule of the invention, comprising culturing the recombinant host cells of the invention.

Also encompassed by the present invention are isolated polypeptides encoded by nucleic acid molecules described herein. For example, the invention relates to an isolated polypeptide comprising an amino acid sequence selected from the group consisting of SEQ ID NOS: 2, 4, 8, 10, 16 and 67-69. The invention also relates to an isolated polypeptide comprising an amino acid sequence having greater than 75 % identity to an amino acid sequence selected from the group consisting of SEQ ID NOS: 2, 4, 8, 10, 16 and 67-69. The invention also provides antibodies, and antigen binding fragments thereof, to the polypeptides of the invention, particularly antibodies and antigen binding fragments thereof which specifically bind the polypeptides described herein.

The invention also provides a method for assaying the presence of a nucleic acid molecule in a sample, comprising contacting said sample with a nucleotide sequence selected from the group consisting of SEQ ID NOS: 1, 3, 7, 9, 11, 12, 13, 14, 15, 17-66, 72 and 73; the complement of SEQ ID NOS: 1, 3, 7, 9, 11, 12, 13, 14, 15, 17-66, 72 and 73; a portion of any one of SEQ ID NOS: 1, 3, 7, 9, 11, 12, 13, 14, 15, 17-66, 72 and 73 which is at least 10 nucleotides in length; and a portion of the complement of any one of SEQ ID NOS: 1, 3, 7, 9, 11, 12, 13, 14, 15, 17-66, 72 and 73 which is at least 10 nucleotides in length, under conditions appropriate for selective hybridization of the sequence to the nucleic acid molecule in the sample. Presence or absence of a hybridization signal indicates presence or absence, respectively, of the target nucleic acid molecule. The invention also relates to a method for assaying the presence of a polypeptide encoded by an isolated nucleic acid molecule of the invention in a sample,

comprising contacting said sample with an antibody which specifically binds to the encoded polypeptide.

The invention further relates to a method of diagnosing or aiding in the diagnosis of neurodegenerative disease in an individual comprising obtaining a nucleic acid sample from the individual and determining the nucleotide present at nucleotide position 5254 of SEQ ID NO: 1, wherein presence of a thymine at said position is indicative of increased likelihood of neurodegenerative disease in the individual as compared with an appropriate control, *e.g.*, an individual having a cytosine at said position. The invention also relates to a method of diagnosing or aiding in the diagnosis of neurodegenerative disease in an individual comprising obtaining a nucleic acid sample from the individual and determining whether there is a deletion of a thymine at nucleotide position 6594 of SEQ ID NO: 1, wherein deletion of a thymine at said position is indicative of increased likelihood of neurodegenerative disease in the individual as compared with an appropriate control, *e.g.*, an individual who does not have a deletion at said position.

The invention also relates to a method of treating a neurodegenerative disorder associated with the presence of a thymine at nucleotide position 5254 of SEQ ID NO: 1 in an individual, comprising administering to the individual an agent selected from the group consisting of a polypeptide encoded by SEQ ID NO: 2 or an active portion thereof, a nucleic acid molecule which encodes SEQ ID NO: 2 or an active portion of SEQ ID NO: 2, and an agonist of SEQ ID NO: 2. The invention further relates to a method of treating a neurodegenerative disorder associated with a deletion at nucleotide position 6594 of SEQ ID NO: 1 in an individual, comprising administering to the individual an agent selected from the group consisting of a polypeptide encoded by SEQ ID NO: 2 or an active portion thereof, a nucleic acid molecule which encodes SEQ ID NO: 2 or an active portion of SEQ ID NO: 2, and an agonist of SEQ ID NO: 2.

The invention also encompasses a method of diagnosing or aiding in the diagnosis of neurodegenerative disease associated with the presence of a thymine at nucleotide position 5254 of SEQ ID NO: 1 in an individual, comprising obtaining a

sample comprising a Spastin polypeptide from the individual and determining the size of the Spastin polypeptide, wherein if the Spastin polypeptide is significantly shorter than SEQ ID NO: 2 it is indicative of neurodegenerative disease. The invention also provides a method of diagnosing or aiding in the diagnosis of neurodegenerative disease associated with the presence of a deletion at nucleotide position 6594 of SEQ ID NO: 1 in an individual, comprising obtaining a sample comprising a Spastin polypeptide from the individual and determining the size of the Spastin polypeptide, wherein if the Spastin polypeptide is significantly shorter than SEQ ID NO: 2 it is indicative of neurodegenerative disease. In one embodiment, the Spastin polypeptide is significantly shorter than SEQ ID NO: 2 if the Spastin polypeptide comprises less than about 75% of the amino acids of SEQ ID NO: 2.

In one embodiment, the neurodegenerative disease comprises one or more symptoms selected from the group consisting of: reduced sensory nerve conduction, reduced motor nerve velocity, hypermyelination of retinal nerve fibers, atrophy of upper cerebellar vermis, absence of Purkinje cells and abnormal neuronal lipid storage. In a particular embodiment, the nucleic acid sample is obtained from a tissue selected from the group consisting of: brain tissue, CNS, lung, fetal lung, testis, lymphocytes, adipose, fibroblasts, skeletal muscle, pancreas, uterus, kidney, tonsil, embryo and isolated cells thereof. For example, brain tissue can be selected from the group consisting of cerebral cortex, granular cell layer of the cerebellum and hippocampus. In a particular embodiment, the neurodegenerative disease is an early onset neurodegenerative disease.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a schematic diagram of the structure and organization of the *spastin* gene. Markers used for the genetic map of the *spastin* gene are shown above. SGCG is the sarcoglycan, gamma gene. hCIT 26_L_1 and hCIT 235_L_20, the overlapping clones that contain the *spastin* ORF, are 110 kilobases (kb) and 60 kb, respectively. Exploded view shows the location of the *spastin* gene. The thick bar is the predicted coding region. The thin bars represent the 5' and 3' UTRs. M is the first methionine. S

is the location fo the introduced stop codon found on the minor haplotype. Δ indicates the location of the deleted base pair found on the major haplotype. AB018273 is the mRNA sequence KIAA0730 (42) which is part of a UniGene cluster (Hs.159492) containing 32 ESTs. R17106, AA776169, AA776670, and AA897178 are additional ESTs with homology to the *spastin* gene.

Figures 2A-2B show the results of sequence analysis and identification of *spastin* mutations found on ARSACS chromosomes. The sequences displayed are from direct sequencing of PCR products and flank the two mutations (indicated by arrows) found on ARSACS chromosomes. Nucleotide numbering is from the putative initiation codon. Figure 2A shows nucleotide 6594 (codon 2198) for an unaffected individual (top panel) and a homozygous affected individual (bottom panel). Figure 2B shows nucleotide 5254 (codon 1752) for an unaffected individual (top panel) and an affected compound heterozygous individual (bottom panel).

Figure 3 shows a Northern blot analysis of *spastin* mRNA. A ^{32}P-labelled 1.8 kb cDNA fragment from the 3' end of the *spastin* gene (Image clone #279258) was hybridized to a blot of fibroblast RNA and to a multiple tissue blot (MTN, Clontech). Lanes 1-5 contain patient fibroblast RNAs and lane 6 contains control fibroblast RNA. The lanes of the MTN blot correspond to the following tissues: 7, heart; 8, brain; 9, placenta; 10, lung; 11, liver; 12, skeletal muscle; 13, kidney; and 14, pancreas. The marker (M) is the 0.24-9.5 kb RNA ladder (Life Technologies).

Figures 4A-4B are schematic representations of the Spastin protein and relevant homologies. Figure 4A shows a schematic representation of the Spastin protein and location of motifs. rep. 1, 2, and 3 represent the domains with homology (28, 30 and 21% identity, respectively) to the *Arabidopsis* open reading frame. Figure 4B shows homology between the two Hsp90 domains of Spastin, the first mouse domain, the *Arabidopsis* open reading frame (GenBank accession #AB006708), and the yeast Hsp90 (GenBank accession #3401959). Alignment was performed with ClustalW (1.7)(43) through the BCM Search Launcher interface (34) with the BLOSUM weight matrix.

The numbering for all sequences is from the first methionine (nucleotide 50,773 is the first methionine of the *Arabidopsis* open reading frame).

Figures 5A-5C show the alignment of the human Spastin with the mouse Spastin. Identical amino acids and gaps are represented by dots and hyphens, respectively. Light gray shading denotes the self-homologous region containing the Hsp90 homology, dark gray shading highlights the DnaJ region. The boxed sequences represent leucine zipper motifs, underlined sequences represent coiled coil domains, and the boxed and underlined sequence delineates the putative hydrophilic region. The first coiled coil domain is interrupted by a proline in the mouse sequence.

Figure 6 is a table showing ESTs identified by sample-sequencing of the ARSACS critical interval.

Figure 7 is a table showing primers for PCR amplification of the human *spastin* gene.

Figures 8A-8G show the complete exon (SEQ ID NO: 3) of the murine *spastin* gene as shown in Figures 8A-8G.

Figures 9A-9F show the complete exon (SEQ ID NO: 1) of the human *spastin* gene.

DETAILED DESCRIPTION OF THE INVENTION

The gene responsible for ARSACS was mapped to chromosome region 13q11 by genotyping 322 microsatellite markers in a genome-wide scan and noting a high degree of homozygosity at locus *D13S787* (Bouchard, J-P., *et al., Neuromuscular Disorders* 8:474-479 (1998)). Extensive genetic analysis of the region defined a maximum multi-point LOD score of 42.3 and revealed a major conserved haplotype among ARSACS chromosomes in a 11.1 cM region flanked by *D13S1236* and *D13S1285* (5). Two groups of ARSACS haplotypes were found between *D13S1275* and *D13S292*. The overwhelming majority (96%) of ARSACS chromosomes carried a single haplotype, defined by *D13S232* and two single nucleotide polymorphisms (SNPs) within the sarcoglycan, gamma gene (SGCG). Location score analysis demonstrated

that the most likely position of the ARSACS was between *D13S232* and *D13S292* (the critical interval)(5).

A high-resolution physical and transcript map of the ARSACS critical interval was constructed in yeast artificial chromosomes (YACs), bacterial artificial chromosomes (BACs) and plasmid artificial chromosomes (PACs). The identification of the ARSACS gene (i.e., a gene in which alteration is associated with ARSACS) was carried out as described herein by performing sample-sequencing of six BAC and PAC clones spanning about 450 kilobases (kb) included in the critical interval. Analysis of the sample sequences revealed human ESTs (Figure 6) and the presence of two known genes: sodium/potassium-ATPase (*ATP1AL1*), that was excluded on the basis of recombination in ARSACS families, and *SGCG*, a gene in which no sequence variants unique to ARSACS chromosomes were found.

A 20 kb sequence contig revealed a huge genomic open reading frame (ORF) of 11,487 base pairs that encodes 3829 amino acids (SEQ ID NO: 2). The open reading frame (ORF) begins with an AUG codon preceded by an in-frame stop codon 75 bp upstream and continues for a total of 3,829 codons before encountering a stop codon. One large cDNA (KIAA0730) derived from a brain library and over 30 ESTs overlap the ORF and allowed the determination of the 3' untranslated region (UTR), which extends 1,307 bp to a polyadenylation site (Figure 1). The existence of this gigantic exon was confirmed by analyzing RT-PCR products spanning the entire mRNA; this analysis showed perfect correspondence between the mRNA and genomic DNA sequence. Thus, the total length of the exon must be at least 12,794 bp. A probe derived from within this sequence detects a transcript of approximately 12.8 kb on a Northern blot, suggesting that the identified exonic sequence may constitute an intronless gene, although the possibility of a small 5' exon cannot be excluded.

To characterize the full sequence of the ORF and to identify potential disease-causing mutations, PCR products from ARSACS patient and control DNA were sequenced. The primers for these reactions are shown in Figure 7. A single-base deletion of a thymine at position 6594 (6594ΔT) (Figure 2A) was found on all copies of

the major ancestral haplotype examined (a total of 32 chromosomes), but was absent in all chromosomes of carrier parents that were not transmitted to ARSACS offspring. This mutation causes a frame shift and results in a subsequent stop codon that truncates the final 43% of the predicted protein. A second mutation, a nonsense mutation of substitution of a thymine for a cytosine at nucleotide position 5254 (c5254T) (Figure 2B) results in the substitution of a stop codon for an arginine and was found on the minor ARSACS haplotype carried in a heterozygous state (in *trans* to the major ARSACS mutation) in six patients from two families (5). Both mutations are thus completely associated with their respective core haplotypes and are predicted to have severe effects on the encoded protein. The presence of these two mutations provides strong evidence that mutations in this ORF are responsible for ARSACS. The gene is referred to herein as *spastin* (gene symbol: SPAS).

In the course of the complete resequencing of the *spastin* gene in ARSACS patients, additional sequence variants were found which proved to be polymorphisms found on non-ARSACS-bearing chromosomes as well. These included four silent substitutions: substitution of a thymine for a cytosine at nucleotide position 3945, substitution of a cytosine for a thymine at nucleotide position 6603, substitution of a thymine for a cytosine at nucleotide position 7731, and substitution of a thymine for a cytosine at nucleotide position 10054 (C3945T, T6603C, C7731T and C10054T, respectively), and an amino acid-altering substitution of a cytosine for a thymine at nucleotide position 7856 (T7856C) that results in the substitution of an alanine for a valine in the predicted protein.

Spastin mRNA was detected by northern blot analysis in fibroblasts, brain and skeletal muscle (Figure 3, lanes 1-6, 8 and 12) and at very low levels in pancreas (Figure 3, lane 14). A single transcript of roughly 12.8 kb was seen in all cases. *Spastin* mRNA was expressed in the fibroblasts of ARSACS patients (Figure 3, lanes 1-5) at the same size as controls, which is not unexpected because both mutations alter only a single nucleotide.

To examine the tissue expression pattern of *spastin* more closely, *in situ* hybridizations were performed. Human, monkey, and rat brain all demonstrated high levels of staining, which included all layers of the cerebral cortex and the granular cell layer of the cerebellum. In a sagittal section of the adult rat brain, strong labeling was seen in most if not all areas of the central nervous system (CNS). Particularly intense labeling was observed on the hippocampus. No labeling is seen with the sense probe. In addition, specific staining of *spastin* mRNA was seen throughout the CNS of the 18-19 day fetal rat. Background staining with the sense probe does not include the CNS. *Spastin* ESTs were identified from the cDNA libraries of many tissues including brain, uterus, kidney, tonsils, liver, and T cells. Transcripts from brain and multiple sclerosis libraries comprise 13 of the 35 human ESTs with homology to *spastin*. Taken together, these lines of evidence indicate that *spastin* is expressed in a variety of tissues, including many that are neural-derived.

On the basis of its amino acid sequence, the Spastin protein product is predicted to have a molecular weight of 437 kD and a pI of 6.85. Structure prediction programs suggest the presence of two leucine zippers, three coiled coils and a hydrophilic domain, all within the C-terminal half of the protein (Figures 4A and 5A-5C). The predicted protein product does not show extensive similarity to any known protein, based on analyses using a variety of different sequence comparison tools. However, two related motifs were identified. The C-terminal portion of the predicted protein contains a 'DnaJ' protein motif (Figures 4A and 5A-5C, residues 3574-3590). Both human and mouse proteins also contain three large segments with sequence similarity to each other, of which two have homology to the N-terminal domain of the Hsp90 class of heat-shock proteins from a variety of organisms. These Hsp90 subdomains are found in *spastin* residues 705-833 and 1773-1895 (Figures 4A and 5A-5C). As discussed below, the DnaJ and Hsp90 protein classes are both involved in molecular chaperone complexes. Interestingly, the three large segments also show strong similarity to a BAC clone recently sequenced as part of the *Arabidopsis* genome project (GenBank #AB006708). Specifically, they are homologous to a portion of a 5,871bp ORF of unknown function

in *Arabidopsis* (Figure 5A). An alignment of the Hsp90 domain is shown for the first and second large segments from human, the first segment for mouse, the *Arabidopsis* ORF and the yeast Hsp90 (Figure 5B). The highly conserved residues correspond to regions already identified as highly conserved "signature sequences" in an extensive phylogenetic analysis of the Hsp90 family (9). Molecular chaperones are known to function in multiple sub-cellular compartments. A knowledge-based program for predicted subcellular localization, PSORT II (10) favored a nuclear localization for the Spastin protein, but the prediction score was relatively weak (47%).

As provided herein, the *spastin* gene is also conserved in mouse. Homologous mouse ESTs were identified, including one having a polyadenylation signal. Using these ESTs to screen a mouse BAC library (CitbCJ7), the mouse *spastin* gene was isolated, identified and sequenced. Sequence analysis of the mouse *spastin* genomic clone revealed the presence of a huge ORF, which is three nucleotides longer than the human homologue and thus, the mouse Spastin protein is predicted to be one amino acid longer. The entire ORF is well conserved between mouse and human, both at the DNA level (88% homology) and at the protein level (94% identity, 97% similarity). The areas of high sequence conservation between mouse and human included the two leucine zippers, two coiled coils, the Hsp90 and DnaJ domains, and the repeated *Arabidopsis* ORF homology (Figures 5A-5C). The 3' UTRs show greater divergence between the mouse and human, but still retain 72% homology. The mouse *spastin* gene was mapped to chromosome 1, near D1Mit373, on the basis of radiation hybrid mapping (LOD score of 25.5) using the Whitehead Institute mouse T31 RH framework (11, 12).

Work described herein strongly supports that a frameshift and a nonsense mutation identified within the *spastin* gene cause ARSACS. Though the gene appears to be widely expressed, the truncation of the Spastin protein caused either by homozygous (6594ΔT/6594ΔT) or compound heterozygous (C5254T/6594ΔT) genotypes apparently lead to symptoms predominantly affecting the nervous system. The high level of expression of *spastin* mRNA in the granular cell layer of the adult rat

INS B6

cerebellum is especially interesting in light of an earlier observation of the reduced thickness of the granular layer found during the postmortem examination of tissue from an ARSACS patient (Bouchard, J-P., *In:* Handbook of Clinical Neurology 16: hereditary neuropathies and spinocerebellar degenerations, pp.451-459, Elsevier Science Publishers, Amsterdam (1991)). Thus, the high mRNA expression levels seen in the CNS indicate a possibly unique role for Spastin in the genesis or maintenance of neural cell function.

~~As described herein, sample sequencing of the ARSACS critical region, in~~ combination with directed sequencing of specific subclones and computer-aided analysis led to the characterization of a very large exon directly from genomic DNA. This likely represents the entire coding sequence of the *spastin* gene as the first methionine is preceded by an in-frame stop codon 75 bp upstream. RT-PCR demonstrated that the sequence, from this 75 bp until the polyadenylation site, is transcribed. *Spastin* appears to be an intronless gene, although a non-coding upstream exon cannot be ruled out. The *spastin* exon of at least 12,794 bp encoding an ORF of 11,487 bp represents the largest exon and the largest ORF within an exon found in any vertebrate so far. The next largest exons reported are the X (inactive)-specific transcript (*XIST*) (11,363 bp) which does not code for a protein (13), and the large central exon of ~~the mucin gene (*MUC5B*) which is 10,713 bp long (14).~~

Intronless ORFs are uncommon and thought to represent at most 5% of human genes. A few gene families are frequently intronless, including histones and G-protein-coupled receptors (GPCRs) (15). Members of the Hsp70 family, but not the Hsp90, are also intronless. The strong conservation between both the human and the mouse *spastin* and the unusually large 5,871 bp *Arabidopsis* ORF suggest both that *spastin* is ancient and that the large size of the exon is functionally important.

The presence of similarities to DnaJ and Hsp90 proteins sheds light on *spastin*'s potential function. Examples of interacting protein pairs having homologues of the two proteins fused into a single protein are well known in the art (17). *Spastin* possesses both the N-terminal domain of the Hsp90 protein class and a DnaJ domain. These two

domains are from proteins that interact in chaperone-mediated protein folding. The DnaJ motif has long been known to form heterocomplexes with the Hsp70 class of proteins in a variety of cellular processes, including ATP-dependent folding of target proteins. The N-terminal domain of the Hsp90 protein class contains an ATP-binding site that is very similar to the one found in DNA gyrase B (18). More recently, it has been shown that the yeast DNAJ homologue, YDJ1, physically associates with Hsp90 and this interaction has specific effects on Hsp90 substrates (19). In addition, other studies have shown that a rabbit DnaJ homologue (p40) interacts with Hsp70 and Hsp90 (both molecular chaperones) to form heterocomplexes known as "foldosomes" (20). Together, these data suggest that *spastin* functions in chaperone-mediated protein folding.

As described herein the mouse *spastin* gene was mapped to chromosome 1 near D1Mit373. A recessive mouse mutation known as tumbler (*tb*: MGI Accession ID:98489) was previously mapped to this region by linkage (21). Tumbler mice had an ataxia that caused them to "walk in a crab-like fashion." They somersaulted, fell over, or jumped when trying to go forward. Most of the homozygotes survived and bred (21). These observations are similar to those seen in ARSACS patients whose life expectancy, although reduced (mean age at death is 51 years) still permits some to survive until the eighth decade. The fertility of affected females seems unchanged, but because overall nuptiality is low, male fertility has been difficult to assess (Bouchard, J-P., *et al., Nueromuscular Disorders* 8:474-479 (1998)). Unfortunately, the *tb* mouse line has died out (Mouse Genome Database: URL:http://www.informatics.jax.org/). However, gene knock-out of the mouse *spastin* gene could serve to confirm that the *tb* mutation was a mutation in the mouse *spastin* gene.

SEQ ID NOS: referred to herein are as follows. SEQ ID NO: 1 refers to the complete exon of the human *spastin* gene as shown in Figures 9A-9F. SEQ ID NO: 2 refers to the protein encoded by the ORF of SEQ ID NO: 1, particularly as shown in Figures 9A-9F and 5A-5C. SEQ ID NO: 3 refers to the complete exon of the murine *spastin* gene as shown in Figures 8A-8G. SEQ ID NO: 4 refers to the protein encoded

by the ORF of SEQ ID NO: 3, particularly as shown in Figures 9A-9F and 5A-5C. SEQ ID NOS: 5 and 6 are intentionally omitted. SEQ ID NO: 7 refers to a nucleotide sequence which is identical to SEQ ID NO: 1 except for a deletion of a thymine at position 6594. SEQ ID NO: 8 refers to the protein encoded by the ORF of SEQ ID NO: 7. SEQ ID NO: 9 refers to a nucleotide sequence which is identical to SEQ ID NO: 1 except for a substitution of a thymine for a cytosine at position 5254. SEQ ID NO: 10 refers to the protein encoded by the ORF of SEQ ID NO: 9. SEQ ID NO: 11, 12, 13 and 14 refer to nucleotide sequences which are identical to SEQ ID NO: 1 except for a substitution of a thymine for a cytosine at position 3945, substitution of a cytosine for a thymine at position 6603, substitution of a thymine for a cytosine at position 7731, and substitution of a thymine for a cytosine at position 10054, respectively. SEQ ID NO: 15 refers to a nucleotide sequence which is identical to SEQ ID NO: 1 except for substitution of a cytosine for a thymine at position 7856. SEQ ID NO: 16 refers to the protein encoded by the ORF of SEQ ID NO: 15. The sequences corresponding to all other SEQ ID NOS: used herein are shown throughout the application.

As appropriate, the isolated nucleic acid molecules of the present invention can be RNA, for example, mRNA, or DNA, such as cDNA and genomic DNA. DNA molecules can be double-stranded or single-stranded; single stranded RNA or DNA can be either the coding, or sense, strand or the non-coding, or antisense, strand. The nucleic acid molecule can include all or a portion of the coding sequence of a gene and can further comprise additional non-coding sequences such as introns and non-coding 3' and 5' sequences (including regulatory sequences, for example). Additionally, the nucleic acid molecule can be fused to a marker sequence, for example, a sequence that encodes a polypeptide to assist in isolation or purification of the polypeptide. Such sequences include, but are not limited to, those which encode a glutathione-S-transferase (GST) fusion protein and those which encode a hemaglutin A (HA) polypeptide marker from influenza.

As used herein, "isolated" is intended to mean that the isolated item is not in the form or environment in which it exists in nature. For example, an "isolated" nucleic

acid molecule, as used herein, is one that is separated from nucleic acid which normally flanks the nucleic acid molecule in nature. With regard to genomic DNA, the term "isolated" refers to nucleic acid molecules which are separated from the chromosome with which the genomic DNA is naturally associated. For example, the isolated nucleic acid molecule can contain less than about 5 kb, 4 kb, 3 kb, 2 kb, 1 kb, 0.5 kb or 0.1 kb of nucleotides which flank the nucleic acid molecule in the genomic DNA of the cell from which the nucleic acid is derived.

Moreover, an isolated nucleic acid of the invention, such as a cDNA or RNA molecule, can be substantially free of other cellular material, or culture medium when produced by recombinant techniques, or chemical precursors or other chemicals when chemically synthesized. However, the nucleic acid molecule can be fused to other coding or regulatory sequences and still be considered isolated. In some instances, the isolated material will form part of a composition (for example, a crude extract containing other substances), buffer system or reagent mix. In other circumstances, the material may be purified to essential homogeneity, for example as determined by PAGE or column chromatography such as HPLC. Preferably, an isolated nucleic acid comprises at least about 50, 80 or 90% (on a molar basis) of all macromolecular species present.

Further, recombinant DNA contained in a vector is included in the definition of "isolated" as used herein. Also, isolated nucleic acid molecules include recombinant DNA molecules in heterologous host cells, as well as partially or substantially purified DNA molecules in solution. "Isolated" nucleic acid molecules also encompass *in vivo* and *in vitro* RNA transcripts of the DNA molecules of the present invention.

The invention further provides variants of the isolated nucleic acid molecules of the invention. Such variants can be naturally occurring, such as allelic variants (same locus), homologs (different locus), and orthologs (different organism), or may be constructed by recombinant DNA methods or by chemical synthesis. Such non-naturally occurring variants can be made using well-known mutagenesis techniques, including those applied to polynucleotides, cells, or organisms.

Accordingly, variants can contain nucleotide substitutions, deletions, inversions and/or insertions in either or both the coding and non-coding region of the nucleic acid molecule. Further, the variations can produce both conservative and non-conservative amino acid substitutions.

Typically, variants have a substantial identity with a nucleic acid molecule disclosed herein and the complements thereof. Particularly preferred are nucleic acid molecules and fragments which have at least about 60%, preferably at least about 70, 80 or 85%, more preferably at least about 90%, even more preferably at least about 95%, and most preferably at least about 98% identity with nucleic acid molecules described herein.

Such nucleic acid molecules can be readily identified as being able to hybridize under stringent conditions to a nucleotide sequence selected from the group consisting of SEQ ID NOS: 1, 3, 7, 9, 11, 12, 13, 14, 15, 17-66, 72 and 73 and the complements thereof. In one embodiment, the variants hybridize under high stringency hybridization conditions (*e.g.*, for selective hybridization) to a nucleotide sequence selected from SEQ ID NOS:1, 3, 7, 9, 11, 12, 13, 14, 15, 17-66, 72 and 73 .

A general description of stringent hybridization conditions are discussed in Ausubel, F.M., *et al.*, *Current Protocols in Molecular Biology*, Greene Publishing Assoc. and Wiley-Interscience 1989, the teachings of which are incorporated herein by reference. Factors such as probe length, base composition, percent mismatch between the hybridizing sequences, temperature and ionic strength influence the stability of nucleic acid hybrids. Thus, stringency conditions sufficient to identify the polynucleotides of the present invention, (e.g., high or moderate stringency conditions) can be determined empirically, depending in part upon the characteristics of the known DNA to which other unknown nucleic acids are being compared for sequence similarity. Equivalent conditions can be determined by varying one or more of these parameters while maintaining a similar degree of identity or similarity between the two nucleic acid molecules. Typically, conditions are used such that sequences at least about 60%, at

least about 70%, at least about 80%, at least about 90% or at least about 95% or more identical to each other remain hybridized to one another.

Alternatively, conditions for stringency are as described in WO 98/40404, the teachings of which are incorporated herein by reference. In particular, examples of highly stringent, stringent, reduced and least stringent conditions are provided in WO 98/40404 in the Table on page 36. In one embodiment, highly stringent conditions are those that are at least as stringent as, for example, 1x SSC at 65°C, or 1x SSC and 50% formamide at 42°C. Moderate stringency conditions are those that are at least as stringent as 4x SSC at 65°C, or 4x SSC and 50% formamide at 42°C. Reduced stringency conditions are those that are at least as stringent as 4x SSC at 50°C, or 6x SSC and 50% formamide at 40°C.

The percent identity of two nucleotide or amino acid sequences can be determined by aligning the sequences for optimal comparison purposes (*e.g.*, gaps can be introduced in the sequence of a first sequence). The nucleotides or amino acids at corresponding positions are then compared, and the percent identity between the two sequences is a function of the number of identical positions shared by the sequences (*i.e.*, % identity = # of identical positions/total # of positions x 100). In certain embodiments, the length of a sequence aligned for comparison purposes is at least 30%, preferably at least 40%, more preferably at least 60%, and even more preferably at least 70%, 80% or 90% of the length of the reference sequence. The actual comparison of the two sequences can be accomplished by well-known methods, for example, using a mathematical algorithm. A preferred, non-limiting example of such a mathematical algorithm is described in Karlin *et al.*, *Proc. Natl. Acad. Sci. USA, 90*:5873-5877 (1993). Such an algorithm is incorporated into the NBLAST and XBLAST programs (version 2.0) as described in Altschul *et al.*, *Nucleic Acids Res., 25*:389-3402 (1997). When utilizing BLAST and Gapped BLAST programs, the default parameters of the respective programs (*e.g.*, NBLAST) can be used. See http://www.ncbi.nlm.nih.gov. In one embodiment, parameters for sequence comparison can be set at score=100, wordlength=12, or can be varied (*e.g.*, W=5 or W=20).

The present invention also provides isolated nucleic acids that contain a fragment or portion that hybridizes under highly stringent conditions to a nucleotide sequence selected from the group consisting of SEQ ID NOS: 1, 3, 7, 9, 11, 12, 13, 14, 15, 17-66, 72 and 73 described herein and the complements of these SEQ ID NOS. The nucleic acid fragments of the invention are at least about 15, preferably at least about 18, 20, 23 or 25 nucleotides, and can be 30, 40, 50, 100, 200 or more nucleotides in length. Longer fragments, for example, 30 or more nucleotides in length, which encode antigenic proteins or polypeptides described herein are useful.

In a related aspect, the nucleic acid fragments of the invention are used as probes or primers in assays such as those described herein. "Probes" are oligonucleotides that hybridize in a base-specific manner to a complementary strand of nucleic acid. Such probes include polypeptide nucleic acids, as described in Nielsen *et al.*, *Science, 254*, 1497-1500 (1991). Typically, a probe comprises a region of nucleotide sequence that hybridizes under highly stringent conditions to at least about 15, typically about 20-25, and more typically about 40, 50 or 75 consecutive nucleotides of a nucleic acid molecule of the invention. More typically, the probe further comprises a label, *e.g.*, radioisotope, fluorescent compound, enzyme, or enzyme co-factor.

As used herein, the term "primer" refers to a single-stranded oligonucleotide which acts as a point of initiation of template-directed DNA synthesis using well-known methods (*e.g.*, PCR, LCR) including, but not limited to those described herein. The appropriate length of the primer depends on the particular use, but typically ranges from about 15 to 30 nucleotides. The term "primer site" refers to the area of the target DNA to which a primer hybridizes. The term "primer pair" refers to a set of primers including a 5' (upstream) primer that hybridizes with the 5' end of the nucleic acid sequence to be amplified and a 3' (downstream) primer that hybridizes with the complement of the sequence to be amplified.

The nucleic acid molecules of the invention such as those described above can be identified and isolated using standard molecular biology techniques and the sequence information provided herein. For example, nucleic acid molecules can be amplified and

isolated by the polymerase chain reaction using synthetic oligonucleotide primers designed based on one or more of the sequences provided herein and the complements thereof. See generally *PCR Technology: Principles and Applications for DNA Amplification* (ed. H.A. Erlich, Freeman Press, NY, NY, 1992); *PCR Protocols: A Guide to Methods and Applications* (Eds. Innis, *et al.*, Academic Press, San Diego, CA, 1990); Mattila *et al.*, *Nucleic Acids Res., 19*:4967 (1991); Eckert *et al.*, *PCR Methods and Applications, 1*:17 (1991); PCR (eds. McPherson *et al.*, IRL Press, Oxford); and U.S. Patent 4,683,202. The nucleic acid molecules can be amplified using cDNA, mRNA or genomic DNA as a template, cloned into an appropriate vector and characterized by DNA sequence analysis.

Other suitable amplification methods include the ligase chain reaction (LCR) (see Wu and Wallace, *Genomics, 4*:560 (1989), Landegren *et al.*, *Science, 241*:1077 (1988), transcription amplification (Kwoh *et al.*, *Proc. Natl. Acad. Sci. USA, 86*:1173 (1989)), and self-sustained sequence replication (Guatelli *et al.*, *Proc. Nat. Acad. Sci. USA, 87*:1874 (1990)) and nucleic acid based sequence amplification (NASBA). The latter two amplification methods involve isothermal reactions based on isothermal transcription, which produce both single stranded RNA (ssRNA) and double stranded DNA (dsDNA) as the amplification products in a ratio of about 30 or 100 to 1, respectively.

For example, the amplified DNA can be radiolabelled and used as a probe for screening a cDNA library derived from fibroblast or brain, *e.g.*, human fibroblast or brain, mRNA in zap express, ZIPLOX or other suitable vector. Corresponding clones can be isolated, DNA can obtained following *in vivo* excision, and the cloned insert can be sequenced in either or both orientations by art recognized methods to identify the correct reading frame encoding a protein of the appropriate molecular weight. For example, the direct analysis of the nucleotide sequence of nucleic acid molecules of the present invention can be accomplished using well-known methods that are commercially available. See, for example, Sambrook *et al.*, *Molecular Cloning, A Laboratory Manual* (2nd Ed., CSHP, New York 1989); Zyskind *et al.*, *Recombinant DNA Laboratory*

Manual, (Acad. Press, 1988)). Using these or similar methods, the protein(s) and the DNA encoding the protein can be isolated, sequenced and further characterized.

Antisense nucleic acids of the invention can be designed using the nucleotide sequences described herein, and constructed using chemical synthesis and enzymatic ligation reactions using procedures known in the art. For example, an antisense nucleic acid (*e.g.*, an antisense oligonucleotide) can be chemically synthesized using naturally occurring nucleotides or variously modified nucleotides designed to increase the biological stability of the molecules or to increase the physical stability of the duplex formed between the antisense and sense nucleic acids, *e.g.*, phosphorothioate derivatives and acridine substituted nucleotides can be used.

In general, the isolated nucleic acid sequences can be used as molecular weight markers on Southern gels, and as chromosome markers which are labeled to map related gene positions. The nucleic acid sequences can also be used to compare with endogenous DNA sequences in patients to identify genetic disorders, and as probes, such as to hybridize and discover related DNA sequences or to subtract out known sequences from a sample. The nucleic acid sequences can further be used to derive primers for genetic fingerprinting, to raise anti-protein antibodies using DNA immunization techniques, and as an antigen to raise anti-DNA antibodies or elicit immune responses. Additionally, the nucleotide sequences of the invention can be used identify and express recombinant proteins for analysis, characterization or therapeutic use, or as markers for tissues in which the corresponding protein is expressed, either constitutively, during tissue differentiation, or in diseased states.

The invention also relates to constructs which comprise a vector into which a sequence of the invention has been inserted in a sense or antisense orientation. As used herein, the term "vector" refers to a nucleic acid molecule capable of transporting another nucleic acid to which it has been linked. One type of vector is a "plasmid", which refers to a circular double stranded DNA loop into which additional DNA segments can be ligated. Another type of vector is a viral vector, wherein additional DNA segments can

be ligated into the viral genome. Certain vectors are capable of autonomous replication in a host cell into which they are introduced (*e.g.*, bacterial vectors having a bacterial origin of replication and episomal mammalian vectors). Other vectors (*e.g.*, non-episomal mammalian vectors) are integrated into the genome of a host cell upon introduction into the host cell, and thereby are replicated along with the host genome. Moreover, certain vectors, expression vectors, are capable of directing the expression of genes to which they are operably linked. In general, expression vectors of utility in recombinant DNA techniques are often in the form of plasmids (vectors). However, the invention is intended to include such other forms of expression vectors, such as viral vectors (*e.g.*, replication defective retroviruses, adenoviruses and adeno-associated viruses) that serve equivalent functions.

Preferred recombinant expression vectors of the invention comprise a nucleic acid of the invention in a form suitable for expression of the nucleic acid in a host cell. This means that the recombinant expression vectors include one or more regulatory sequences, selected on the basis of the host cells to be used for expression, which is operably linked to the nucleic acid sequence to be expressed. Within a recombinant expression vector, "operably linked" is intended to mean that the nucleotide sequence of interest is linked to the regulatory sequence(s) in a manner which allows for expression of the nucleotide sequence (*e.g.*, in an *in vitro* transcription/translation system or in a host cell when the vector is introduced into the host cell). The term "regulatory sequence" is intended to include promoters, enhancers and other expression control elements (*e.g.*, polyadenylation signals). Such regulatory sequences are described, for example, in Goeddel, *Gene Expression Technology: Methods in Enzymology 185*, Academic Press, San Diego, CA (1990). Regulatory sequences include those which direct constitutive expression of a nucleotide sequence in many types of host cell and those which direct expression of the nucleotide sequence only in certain host cells (*e.g.*, tissue-specific regulatory sequences). It will be appreciated by those skilled in the art that the design of the expression vector can depend on such factors as the choice of the host cell to be transformed, the level of expression of protein desired, etc.

The expression vectors of the invention can be introduced into host cells to thereby produce proteins or peptides, including fusion proteins or peptides, encoded by nucleic acids as described herein . The recombinant expression vectors of the invention can be designed for expression of a polypeptide of the invention in prokaryotic or eukaryotic cells, *e.g.*, bacterial cells such as *E. coli*, insect cells (using baculovirus expression vectors), yeast cells or mammalian cells. Suitable host cells are discussed further in Goeddel, *supra*. Alternatively, the recombinant expression vector can be transcribed and translated *in vitro*, for example using T7 promoter regulatory sequences and T7 polymerase.

Another aspect of the invention pertains to host cells into which a recombinant expression vector of the invention has been introduced. The terms "host cell" and "recombinant host cell" are used interchangeably herein. It is understood that such terms refer not only to the particular subject cell but also to the progeny or potential progeny of such a cell. Because certain modifications may occur in succeeding generations due to either mutation or environmental influences, such progeny may not, in fact, be identical to the parent cell, but are still included within the scope of the term as used herein.

A host cell can be any prokaryotic or eukaryotic cell. For example, a nucleic acid of the invention can be expressed in bacterial cells (*e.g.*, *E. coli*), insect cells, yeast or mammalian cells (such as Chinese hamster ovary cells (CHO) or COS cells). Other suitable host cells are known to those skilled in the art.

Vector DNA can be introduced into prokaryotic or eukaryotic cells via conventional transformation or transfection techniques. As used herein, the terms "transformation" and "transfection" are intended to refer to a variety of art-recognized techniques for introducing foreign nucleic acid (*e.g.*, DNA) into a host cell, including calcium phosphate or calcium chloride co-precipitation, DEAE-dextran-mediated transfection, lipofection, or electroporation. Suitable methods for transforming or transfecting host cells can be found in Sambrook, *et al.* (*supra*), and other laboratory manuals.

A host cell of the invention, such as a prokaryotic or eukaryotic host cell in culture, can be used to produce (*i.e.*, express) a polypeptide of the invention. Accordingly, the invention further provides methods for producing a polypeptide using the host cells of the invention. In one embodiment, the method comprises culturing the host cell of invention (into which a recombinant expression vector encoding a polypeptide of the invention has been introduced) in a suitable medium such that the polypeptide is produced. In another embodiment, the method further comprises isolating the polypeptide from the medium or the host cell.

The host cells of the invention can also be used to produce nonhuman transgenic animals. For example, in one embodiment, a host cell of the invention is a fertilized oocyte or an embryonic stem cell into which a nucleic acid of the invention have been introduced. Such host cells can then be used to create non-human transgenic animals in which exogenous nucleotide sequences have been introduced into their genome or homologous recombinant animals in which endogenous nucleotide sequences have been altered. Such animals are useful for studying the function and/or activity of the nucleotide sequence and polypeptide encoded by the sequence and for identifying and/or evaluating modulators of their activity. As used herein, a "transgenic animal" is a non-human animal, preferably a mammal, more preferably a rodent such as a rat or mouse, in which one or more of the cells of the animal includes a transgene. Other examples of transgenic animals include non-human primates, sheep, dogs, cows, goats, chickens, amphibians, etc. A transgene is exogenous DNA which is integrated into the genome of a cell from which a transgenic animal develops and which remains in the genome of the mature animal, thereby directing the expression of an encoded gene product in one or more cell types or tissues of the transgenic animal. As used herein, an "homologous recombinant animal" is a non-human animal, preferably a mammal, more preferably a mouse, in which an endogenous gene has been altered by homologous recombination between the endogenous gene and an exogenous DNA molecule introduced into a cell of the animal, *e.g.*, an embryonic cell of the animal, prior to development of the animal.

A transgenic animal of the invention can be created by introducing a nucleic acid of the invention into the male pronuclei of a fertilized oocyte, *e.g.*, by microinjection, retroviral infection, and allowing the oocyte to develop in a pseudopregnant female foster animal. The sequence can be introduced as a transgene into the genome of a non-human animal. Intronic sequences and polyadenylation signals can also be included in the transgene to increase the efficiency of expression of the transgene. A tissue-specific regulatory sequence(s) can be operably linked to the transgene to direct expression of a polypeptide in particular cells. Methods for generating transgenic animals via embryo manipulation and microinjection, particularly animals such as mice, have become conventional in the art and are described, for example, in U.S. Patent Nos. 4,736,866 and 4,870,009, U.S. Patent No. 4,873,191 and in Hogan, *Manipulating the Mouse Embryo* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 1986). Similar methods are used for production of other transgenic animals. A transgenic founder animal can be identified based upon the presence of the transgene in its genome and/or expression of mRNA in tissues or cells of the animals. A transgenic founder animal can then be used to breed additional animals carrying the transgene. Moreover, transgenic animals carrying a transgene encoding the transgene can further be bred to other transgenic animals carrying other transgenes.

The present invention also provides isolated polypeptides and variants and fragments thereof that are encoded by the nucleic acid molecules of the invention. For example, as described above, the nucleotide sequences can be used to design primers to clone and express cDNAs encoding the polypeptides of the invention.

As used herein, a polypeptide is said to be "isolated" or "purified" when it is substantially free of cellular material when it is isolated from recombinant and non-recombinant cells, or free of chemical precursors or other chemicals when it is chemically synthesized. A polypeptide, however, can be joined to another polypeptide with which it is not normally associated in a cell and still be "isolated" or "purified."

The polypeptides of the invention can be purified to homogeneity. It is understood, however, that preparations in which the polypeptide is not purified to

homogeneity are useful and considered to contain an isolated form of the polypeptide. The critical feature is that the preparation allows for the desired function of the polypeptide, even in the presence of considerable amounts of other components. Thus, the invention encompasses various degrees of purity. In one embodiment, the language "substantially free of cellular material" includes preparations of the polypeptide having less than about 30% (by dry weight) other proteins (*i.e.*, contaminating protein), less than about 20% other proteins, less than about 10% other proteins, or less than about 5% other proteins.

When a polypeptide is recombinantly produced, it can also be substantially free of culture medium, *i.e.*, culture medium represents less than about 20%, less than about 10%, or less than about 5% of the volume of the protein preparation. The language "substantially free of chemical precursors or other chemicals" includes preparations of the polypeptide in which it is separated from chemical precursors or other chemicals that are involved in its synthesis. In one embodiment, the language "substantially free of chemical precursors or other chemicals" includes preparations of the polypeptide having less than about 30% (by dry weight) chemical precursors or other chemicals, less than about 20% chemical precursors or other chemicals, less than about 10% chemical precursors or other chemicals, or less than about 5% chemical precursors or other chemicals.

In one embodiment, a polypeptide comprises an amino acid sequence selected from the group consisting of SEQ ID NOS: 2, 4, 8, 10 and 16 and the complements thereof. However, the invention also encompasses sequence variants. Variants include a substantially homologous protein encoded by the same genetic locus in an organism, *i.e.*, an allelic variant. Variants also encompass proteins derived from other genetic loci in an organism, but having substantial homology to a polypeptide of the invention. Variants also include proteins substantially homologous to these polypeptides but derived from another organism, *i.e.*, an ortholog. Variants also include proteins that are substantially homologous to these polypeptides that are produced by chemical synthesis. Variants also

include proteins that are substantially homologous or identical to these polypeptides that are produced by recombinant methods.

As used herein, two proteins (or a region of the proteins) are substantially homologous or identical when the amino acid sequences are at least about 45-55%, typically at least about 70-75%, more typically at least about 80-85%, and most typically at least about 90-95% or more homologous or identical. A substantially homologous amino acid sequence, according to the present invention, will be encoded by a nucleic acid hybridizing to a nucleic acid sequence described herein, or portion thereof, under stringent conditions as more described above.

To determine the percent homology or identity of two amino acid sequences, or of two nucleic acids, the sequences are aligned for optimal comparison purposes (*e.g.*, gaps can be introduced in the sequence of one protein or nucleic acid for optimal alignment with the other protein or nucleic acid). The amino acid residues or nucleotides at corresponding amino acid positions or nucleotide positions are then compared. When a position in one sequence is occupied by the same amino acid residue or nucleotide as the corresponding position in the other sequence, then the molecules are homologous at that position. As used herein, amino acid or nucleic acid "homology" is equivalent to amino acid or nucleic acid "identity". The percent homology between the two sequences is a function of the number of identical positions shared by the sequences (*i.e.*, per cent homology equals the number of identical positions/total number of positions times 100).

The invention also encompasses polypeptides having a lower degree of identity but having sufficient similarity so as to perform one or more of the same functions performed by a polypeptide encoded by a nucleic acid of the invention. Similarity is determined by conserved amino acid substitution. Such substitutions are those that substitute a given amino acid in a polypeptide by another amino acid of like characteristics. Conservative substitutions are likely to be phenotypically silent. Typically seen as conservative substitutions are the replacements, one for another, among the aliphatic amino acids Ala, Val, Leu, and Ile; interchange of the hydroxyl residues Ser and Thr, exchange of the acidic residues Asp and Glu, substitution between the amide

residues Asn and Gln, exchange of the basic residues Lys and Arg and replacements among the aromatic residues Phe, Tyr. Guidance concerning which amino acid changes are likely to be phenotypically silent are found in Bowie *et al.*, Science 247:1306-1310 (1990).

Preferred computer program methods to determine identify and similarity between two sequences include, but are not limited to, GCG program package (Devereux, J., *et al.*, *Nucleic Acids Res.*, *12*(1):387 (1984)), BLASTP, BLASTN, FASTA (Atschul, S.F. *et al.*, *J. Molec. Biol.*, *215*:403 (1990)).

A variant polypeptide can differ in amino acid sequence by one or more substitutions, deletions, insertions, inversions, fusions, and truncations or a combination of any of these. Further, variant polypeptides can be fully functional or can lack function in one or more activities. Fully functional variants typically contain only conservative variation or variation in non-critical residues or in non-critical regions. Functional variants can also contain substitution of similar amino acids that result in no change or an insignificant change in function. Alternatively, such substitutions may positively or negatively affect function to some degree.

Non-functional variants typically contain one or more non-conservative amino acid substitutions, deletions, insertions, inversions, or truncation or a substitution, insertion, inversion, or deletion in a critical residue or critical region. As indicated, variants can be naturally-occurring or can be made by recombinant means or chemical synthesis to provide useful and novel characteristics for the polypeptide. This includes preventing immunogenicity from pharmaceutical formulations by preventing protein aggregation.

Amino acids that are essential for function can be identified by methods known in the art, such as site-directed mutagenesis or alanine-scanning mutagenesis (Cunningham *et al.*, *Science, 244*:1081-1085 (1989)). The latter procedure introduces single alanine mutations at every residue in the molecule. The resulting mutant molecules are then tested for biological activity *in vitro*, or *in vitro* proliferative activity. Sites that are critical for polypeptide activity can also be determined by structural analysis such as

crystallization, nuclear magnetic resonance or photoaffinity labeling (Smith *et al.*, *J. Mol. Biol., 224*:899-904 (1992); de Vos *et al. Science, 255*:306-312 (1992)).

The invention also includes polypeptide fragments or portions of the polypeptides of the invention, as well as fragments of the variants of the polypeptides described herein. As used herein, a fragment comprises at least 6 contiguous amino acids. Useful fragments include those that retain one or more of the biological activities of the polypeptide as well as fragments that can be used as an immunogen to generate polypeptide specific antibodies.

Biologically active fragments (peptides which are, for example, 6, 9, 12, 15, 20, 30, 35, 36, 37, 38, 39, 40, 50, 100 or more amino acids in length) can comprise a domain, segment, or motif that has been identified by analysis of the polypeptide sequence using well-known methods, *e.g.*, signal peptides, extracellular domains, one or more transmembrane segments or loops, ligand binding regions, zinc finger domains, DNA binding domains, acylation sites, glycosylation sites, or phosphorylation sites.

The invention also provides fragments with immunogenic properties. These contain an epitope-bearing portion of the polypeptides and variants of the invention. These epitope-bearing peptides are useful to raise antibodies that bind specifically to a polypeptide or region or fragment. These peptides can contain at least 6, 7, 8, 9, 12, at least 14, or between at least about 15 to about 30 amino acids. The epitope-bearing peptide and polypeptides may be produced by any conventional means (Houghten, R.A., *Proc. Natl. Acad. Sci. USA, 82*:5131-5135 (1985)). Simultaneous multiple peptide synthesis is described in U.S. Patent No. 4,631,211.

Fragments can be discrete (not fused to other amino acids or polypeptides) or can be within a larger polypeptide. Further, several fragments can be comprised within a single larger polypeptide. In one embodiment a fragment designed for expression in a host can have heterologous pre- and pro-polypeptide regions fused to the amino terminus of the polypeptide fragment and an additional region fused to the carboxyl terminus of the fragment.

The invention thus provides chimeric or fusion proteins. These comprise a polypeptide of the invention operatively linked to a heterologous protein having an amino acid sequence not substantially homologous to the polypeptide. "Operatively linked" indicates that the polypeptide protein and the heterologous protein are fused in-frame. The heterologous protein can be fused to the N-terminus or C-terminus of the polypeptide. In one embodiment the fusion protein does not affect function of the polypeptide *per se*. For example, the fusion protein can be a GST-fusion protein in which the polypeptide sequences are fused to the C-terminus of the GST sequences. The isolated polypeptide can be purified from cells that naturally express it, such as from mammary epithelium, purified from cells that have been altered to express it (recombinant), or synthesized using known protein synthesis methods.

In one embodiment, the protein is produced by recombinant DNA techniques. For example, a nucleic acid molecule encoding the polypeptide is cloned into an expression vector, the expression vector introduced into a host cell and the protein expressed in the host cell. The protein can then be isolated from the cells by an appropriate purification scheme using standard protein purification techniques.

Polypeptides often contain amino acids other than the 20 amino acids commonly referred to as the 20 naturally-occurring amino acids. Further, many amino acids, including the terminal amino acids, may be modified by natural processes, such as processing and other post-translational modifications, or by chemical modification techniques well known in the art. Common modifications that occur naturally in polypeptides are described in basic texts, detailed monographs, and the research literature, and they are well known to those of skill in the art.

Accordingly, the polypeptides also encompass derivatives or analogs in which a substituted amino acid residue is not one encoded by the genetic code, in which a substituent group is included, in which the mature polypeptide is fused with another compound, such as a compound to increase the half-life of the polypeptide (for example, polyethylene glycol), or in which the additional amino acids are fused to the mature

polypeptide, such as a leader or secretory sequence or a sequence for purification of the mature polypeptide or a pro-protein sequence.

In general, polypeptides or proteins of the present invention can be used as a molecular weight marker on SDS-PAGE gels or on molecular sieve gel filtration columns using art-recognized methods. The polypeptides of the present invention can be used to raise antibodies or to elicit an immune response. The polypeptides can also be used as a reagent, *e.g.*, a labeled reagent, in assays to quantitatively determine levels of the protein or a molecule to which it binds (*e.g.*, a receptor or a ligand) in biological fluids. The polypeptides can also be used as markers for tissues in which the corresponding protein is preferentially expressed, either constitutively, during tissue differentiation, or in a diseased state. The polypeptides can be used to isolate a corresponding binding partner, *e.g.*, receptor or ligand, such as, for example, in an interaction trap assay, and to screen for peptide or small molecule antagonists or agonists of the binding interaction.

In another aspect, the invention provides antibodies to the polypeptides and polypeptide fragments of the invention. The term "antibody" as used herein refers to immunoglobulin molecules and immunologically active portions of immunoglobulin molecules, *i.e.*, molecules that contain an antigen binding site that specifically binds an antigen. A molecule that specifically binds to a polypeptide of the invention is a molecule that binds to that polypeptide or a fragment thereof, but does not substantially bind other molecules in a sample, *e.g.*, a biological sample, which naturally contains the polypeptide. Examples of immunologically active portions of immunoglobulin molecules include F(ab) and $F(ab')_2$ fragments which can be generated by treating the antibody with an enzyme such as pepsin. The invention provides polyclonal and monoclonal antibodies that bind to a polypeptide of the invention; such antibodies can be made using methods known in the art. The term "monoclonal antibody" or "monoclonal antibody composition", as used herein, refers to a population of antibody molecules that contain only one species of an antigen binding site capable of immunoreacting with a particular epitope of a polypeptide of the invention. A monoclonal antibody composition

thus typically displays a single binding affinity for a particular polypeptide of the invention with which it immunoreacts.

Additionally, recombinant antibodies, such as chimeric and humanized monoclonal antibodies, comprising both human and non-human portions, which can be made using standard recombinant DNA techniques, are within the scope of the invention. Such chimeric and humanized monoclonal antibodies can be produced by recombinant DNA techniques known in the art, for example using methods described in PCT Publication No. WO 87/02671; European Patent Application 184,187; European Patent Application 171,496; European Patent Application 173,494; PCT Publication No. WO 86/01533; U.S. Patent No. 4,816,567; European Patent Application 125,023; Better *et al.* (1988) *Science, 240*:1041-1043; Liu *et al.* (1987) *Proc. Natl. Acad. Sci. USA, 84*:3439-3443; Liu *et al.* (1987) *J. Immunol., 139*:3521-3526; Sun *et al.* (1987) *Proc. Natl. Acad. Sci. USA, 84*:214-218; Nishimura *et al.* (1987) *Canc. Res., 47*:999-1005; Wood *et al.* (1985) *Nature, 314*:446-449; and Shaw *et al.* (1988) *J. Natl. Cancer Inst., 80*:1553-1559); Morrison (1985) *Science, 229*:1202-1207; Oi *et al.* (1986) *Bio/Techniques, 4*:214; U.S. Patent 5,225,539; Jones *et al.* (1986) *Nature, 321*:552-525; Verhoeyan *et al.* (1988) *Science, 239*:1534; and Beidler *et al.* (1988) *J. Immunol., 141*:4053-4060.

In general, antibodies of the invention (*e.g.*, a monoclonal antibody) can be used to isolate a polypeptide of the invention by standard techniques, such as affinity chromatography or immunoprecipitation. A polypeptide specific antibody can facilitate the purification of natural polypeptide from cells and of recombinantly produced polypeptide expressed in host cells. Moreover, an antibody specific for a polypeptide of the invention can be used to detect the polypeptide (*e.g.*, in a cellular lysate, cell supernatant, or tissue sample) in order to evaluate the abundance and pattern of expression of the polypeptide. Antibodies can be used diagnostically to monitor protein levels in tissue as part of a clinical testing procedure, *e.g.*, to, for example, determine the efficacy of a given treatment regimen. Detection can be facilitated by coupling the antibody to a detectable substance. Examples of detectable substances include various

enzymes, prosthetic groups, fluorescent materials, luminescent materials, bioluminescent materials, and radioactive materials. Examples of suitable enzymes include horseradish peroxidase, alkaline phosphatase, (β-galactosidase, or acetylcholinesterase; examples of suitable prosthetic group complexes include streptavidin/biotin and avidin/biotin; examples of suitable fluorescent materials include umbelliferone, fluorescein, fluorescein isothiocyanate, rhodamine, dichlorotriazinylamine fluorescein, dansyl chloride or phycoerythrin; an example of a luminescent material includes luminol; examples of bioluminescent materials include luciferase, luciferin, and aequorin, and examples of suitable radioactive material include ^{125}I, ^{131}I, ^{35}S or ^{3}H.

The present invention also pertains to diagnostic assays and prognostic assays used for prognostic (predictive) purposes to thereby treat an individual prophylactically. Accordingly, one aspect of the present invention relates to diagnostic assays for determining protein and/or nucleic acid expression as well as activity of proteins of the invention, in the context of a biological sample (*e.g.*, blood, serum, cells, tissue) to thereby determine whether an individual is afflicted with a disease or disorder, or is at risk of developing a disorder, e.g., a neurodegenerative disorders such as ARSACS, associated with aberrant expression or activity. The invention also provides for prognostic (or predictive) assays for determining whether an individual is at risk of developing a disorder associated with activity or expression of proteins or nucleic acids of the invention.

Disorders which may be treated or diagnosed by methods described herein include, but are not limited to, neurodegenerative disease comprising one or more symptoms or effects selected from the group consisting of: reduced sensory nerve conduction, reduced motor nerve velocity, hypermyelination of retinal nerve fibers, atrophy of upper cerebellar vermis, absence of Purkinje cells and abnormal neuronal lipid storage. The invention is particularly suited to treat and diagnose ARSACS.

Another aspect of the invention pertains to monitoring the influence of agents (*e.g.*, drugs, compounds) on the expression or activity of proteins of the invention in clinical trials.

An exemplary method for detecting the presence or absence of proteins or nucleic acids of the invention in a biological sample involves obtaining a biological sample from a test subject and contacting the biological sample with a compound or an agent capable of detecting the protein, or nucleic acid (*e.g.*, mRNA, genomic DNA) that encodes the protein, such that the presence of the protein or nucleic acid is detected in the biological sample. A preferred agent for detecting mRNA or genomic DNA is a labeled nucleic acid probe capable of hybridizing to mRNA or genomic DNA sequences described herein. The nucleic acid probe can be, for example, a full-length nucleic acid, or a portion thereof, such as an oligonucleotide of at least 15, 30, 50, 100, 250 or 500 nucleotides in length and sufficient to specifically hybridize under stringent conditions to appropriate mRNA or genomic DNA. Other suitable probes for use in the diagnostic assays of the invention are described herein.

In one embodiment, the agent for detecting proteins of the invention is an antibody capable of binding to the protein, preferably an antibody with a detectable label. Antibodies can be polyclonal, or more preferably, monoclonal. An intact antibody, or a fragment thereof (*e.g.*, Fab or $F(ab')_2$) can be used. The term "labeled", with regard to the probe or antibody, is intended to encompass direct labeling of the probe or antibody by coupling (*i.e.*, physically linking) a detectable substance to the probe or antibody, as well as indirect labeling of the probe or antibody by reactivity with another reagent that is directly labeled. Examples of indirect labeling include detection of a primary antibody using a fluorescently labeled secondary antibody and end-labeling of a DNA probe with biotin such that it can be detected with fluorescently labeled streptavidin. In a preferred embodiment, the antibody is able to distinguish between complete or nearly complete proteins and truncated versions of the same protein.

The term "biological sample" is intended to include tissues, calls and biological fluids isolated from a subject, as well as tissues, cells and fluids present within a subject. For example, the sample can be obtained from a tissue selected from the group consisting of: brain tissue, CNS, lung, fetal lung, testis, lymphocytes, adipose, fibroblasts, skeletal muscle, pancreas, uterus, kidney, tonsil, embryo and isolated cells thereof. That is, the

detection method of the invention can be used to detect mRNA, protein, or genomic DNA of the invention in a biological sample *in vitro* as well as *in vivo*. For example, *in vitro* techniques for detection of mRNA include Northern hybridizations and *in situ* hybridizations. *In vitro* techniques for detection of protein include enzyme linked immunosorbent assays (ELISAs), Western blots, immunoprecipitations and immunofluorescence. *In vitro* techniques for detection of genomic DNA include Southern hybridizations. Furthermore, *in vivo* techniques for detection of protein include introducing into a subject a labeled anti-protein antibody. For example, the antibody can be labeled with a radioactive marker whose presence and location in a subject can be detected by standard imaging techniques.

In one embodiment, the biological sample contains protein molecules from the test subject. Alternatively, the biological sample can contain mRNA molecules from the test subject or genomic DNA molecules from the test subject. A preferred biological sample is a serum sample or mammary epithelium isolated by conventional means from a subject. A nucleic acid sample is a sample, *e.g.*, a biological sample, which contains nucleic acid molecules.

The invention also encompasses kits for detecting the presence of proteins or nucleic acid molecules of the invention in a biological sample. For example, the kit can comprise a labeled compound or agent capable of detecting protein or mRNA in a biological sample; means for determining the amount of in the sample; and means for comparing the amount of in the sample with a standard. The compound or agent can be packaged in a suitable container. The kit can further comprise instructions for using the kit to detect protein or nucleic acid.

The diagnostic methods described herein can also be utilized to identify subjects having or at risk of developing a disease or disorder associated with aberrant expression or activity of proteins and nucleic acid molecules of the invention. For example, the assays described herein can be utilized to identify a subject having or at risk of developing a disorder associated with Spastin protein or *spastin* nucleic acid expression or activity such as a neurodegenerative disorder. Thus, the present invention provides a

method for identifying a disease or disorder associated with aberrant expression or activity of proteins or nucleic acid molecules of the invention, in which a test sample is obtained from a subject and protein or nucleic acid molecule (*e.g.*, mRNA, genomic DNA) is detected, wherein the presence of an altered protein or nucleic acid molecule is diagnostic for a subject having or at risk of developing a disease or disorder associated with aberrant expression or activity of the protein or nucleic acid sequence of the invention. In certain embodiments as described herein, it is valuable to determine the genotype of an individual, particularly where a specific allelic form is associated with disease. For example, it will be valuable for purposes of diagnosis to determine an individual's genotype for the C52454T mutation with respect to ARSACS diagnosis, *i.e.*, to identify alteration in the *spastin* gene or Spastin protein.

Detection of the alteration can involve the use of a probe/primer in a polymerase chain reaction (PCR) (see, *e.g.*, U.S. Patent Nos. 4,683,195 and 4,683,202), such an anchor PCR or RACE PCR, or, alternatively, in a ligation chain reaction (LCR) (see, *e.g.*, Landegran *et al.* (1988) *Science, 241*:1077-1080; and Nakazawa *et al.* (1994) *PNAS, 91*:360-364), the latter of which can be particularly useful for detecting point mutations (see Abravaya *et al.* (1995) *Nucleic Acids Res., 23*:675-682). This method can include the steps of collecting a sample of cells from a patient, isolating nucleic acid molecules (*e.g.*, genomic, mRNA or both) from the cells of the sample, contacting the nucleic acid sample with one or more primers which specifically hybridize to the gene under conditions such that hybridization and amplification of the gene (if present) occurs, and detecting the presence or absence of an amplification product, or detecting the size of the amplification product and comparing the length to a control sample. It is anticipated that PCR and/or LCR may be desirable to use as a preliminary amplification step in conjunction with any of the techniques used for detecting mutations described herein. In one embodiment allele-specific primers are utilized.

Alternative amplification methods include: self sustained sequence replication (Guatelli, J.C. *et al.* (1990) *Proc. Natl. Acad. Sci. USA, 87*:1874-1878), transcriptional amplification system (Kwoh, D.Y. *et al.*, (1989) *Proc. Natl. Acad. Sci. USA,*

86:1173-1177), Q-Beta Replicase (Lizardi, P.M. *et al.*,(1988) *Bio/Technology, 6*:1197), or any other nucleic acid amplification method, followed by the detection of the amplified molecules using techniques well known to those of skill in the art. These detection schemes are especially useful for the detection of nucleic acid molecules if such molecules are present in very low numbers.

In an alternative embodiment, mutations in a given gene from a sample cell can be identified by alterations in restriction enzyme cleavage patterns. For example, sample and control DNA is isolated, amplified (optionally), digested with one or more restriction endonucleases, and fragment length sizes are determined by gel electrophoresis and compared. Differences in fragment length sizes between sample and control DNA indicate mutations in the sample DNA. Moreover, the use of sequence specific ribozymes (see, for sample, U.S. Patent No. 5,498,531) can be used to score for the presence of specific mutations, *e.g.*, the C5254T mutation, by development or loss of a ribozyme cleavage site.

In other embodiments, genetic mutations can be identified by hybridizing a sample and control nucleic acids, *e.g.*, DNA or RNA, to high density arrays containing many oligonucleotide probes (Cronin, M.T. *et al.* (1996) *Human Mutation, 7*:244-255; Kozal, M.J. *et al.*(1996) *Nature Medicine, 2*:753-759). For example, genetic mutations can be identified in two dimensional arrays containing light-generated DNA probes as described in Cronin, M.T. *et al. supra*. Briefly, a first hybridization array of probes can be used to scan through long stretches of DNA in a sample and control to identify base changes between the sequences by making linear arrays of sequential overlapping probes. This step allows the identification of point mutations. This step is followed by a second hybridization array that allows the characterization of specific mutations by using smaller, specialized probe arrays complementary to all variants or mutations detected. Each mutation array is composed of parallel probe sets, one complementary to the wild-type gene and the other complementary to the mutant gene.

In yet another embodiment, any of a variety of sequencing reactions known in the art can be used to directly sequence the gene and detect specific mutations by comparing

the sequence of the gene from the sample with the corresponding wild-type (control) gene sequence. Examples of sequencing reactions include those based on techniques developed by Maxim and Gilbert ((1997) *PNAS, 74*:560) or Sanger ((1977) *PNAS, 74*:5463). It is also contemplated that any of a variety of automated sequencing procedures can be utilized when performing the diagnostic assays ((1995) *Biotechniques, 19*:448), including sequencing by mass spectrometry (see, *e.g.*, PCT International Publication No. WO 94/16101; Cohen *et al.* (1996) *Adv. Chromatogr., 36*:127-162; and Griffin *et al.* (1993) *Appl. Biochem. Biotechnol., 38*:147-159).

Other methods for detecting mutations include methods in which protection from cleavage agents is used to detect mismatched bases in RNA/RNA or RNA/DNA heteroduplexes (Myers *et al.* (1985) *Science, 230*:1242). In general, the art technique of "mismatch cleavage" starts by providing heteroduplexes formed by hybridizing (labeled) RNA or DNA containing the wild-type sequence with potentially mutant RNA or DNA obtained from a tissue sample. The double-standard duplexes are treated with an agent that cleaves single-stranded regions of the duplex such as which will exist due to base pair mismatches between the control and sample strands. For instance, RNA/DNA duplexes can be treated with Rnase and DNA/DNA hybrids treated with S1 nuclease to enzymatically digest the mismatched regions. After digestion of the mismatched regions, the resulting material is then separated by size on denaturing polyacrylamide gels to determine the site of mutation. See, for example Cotton *et al.* (1988) *Proc. Natl. Acad. Sci. USA, 85*:4397; Saleeba *et al.* (1992) *Methods Enzymol., 217*:286-295. In certain embodiments, the control DNA or RNA can be labeled for detection.

In still another embodiment, the mismatch cleavage reaction employs one or more proteins that recognize mismatched base pairs in double-stranded DNA (so called "DNA mismatch repair" enzymes) in defined systems for detecting and mapping point mutations in cDNAs obtained from samples of cells. For example, the mutY enzyme of *E. coli* cleaves A at G/A mismatches and the thymidine DNA glycosylase from HeLa cells cleaves T at G/T mismatches (Hsu *et al.* (1994) *Carcinogenesis, 15*:1657-1662). According to an exemplary embodiment, a probe based on an nucleotide sequence of the

invention is hybridized to a cDNA or other DNA product from a test cell(s). The duplex is treated with a DNA mismatch repair enzyme, and the cleavage products, if any, can be detected from electrophoresis protocols or the like. See, for example, U.S. Patent No. 5,459,039.

In other embodiments, alterations in electrophoretic mobility will be used to identify mutations in nucleic acid molecules described herein. For example, single strand conformation polymorphism (SSCP) may be used to detect differences in electrophoretic mobility between mutant and wild type nucleic acids (Orita *et al.* (1989) Proc. *Natl. Acad. Sci. USA, 86*:2766, see also Cotton (1993) *Mutat Res, 285*:125-144; and Hayashi (1992) *Genet Anal. Tech. Appl., 9*:73-79). Single-stranded DNA fragments of sample and control nucleic acids will be denatured and allowed to renature. The secondary structure of single-stranded nucleic acids varies according to sequence, and the resulting alteration in electrophoretic mobility enables the detection of even a single base change. The DNA fragments may be labeled or detected with labeled probes. The sensitivity of the assay may be enhanced by using RNA (rather than DNA), in which the secondary structure is more sensitive to a change in sequence. In one embodiment, the subject method utilizes heteroduplex analysis to separate double stranded heteroduplex molecules on the basis of changes in electrophoretic mobility (Keen *et al.* (1991) *Trends Genet., 7*:5).

In yet another embodiment the movement of mutant or wild-type fragments in polyacrylamide gels containing a gradient of denaturant is assayed using denaturing gradient gel electrophoresis (DGGE) (Myers *et al.* (1985) *Nature, 313*:495). When DGGE is used as the method of analysis, DNA will be modified to insure that it does not completely denature, for example by adding a GC clamp of approximately 40 bp of high-melting GC-rich DNA by PCR. In a further embodiment, a temperature gradient is used in place of a denaturing gradient to identify differences in the mobility of control and sample DNA (Rosenbaum and Reissner (1987) *Biophys. Chem., 265*:12753).

Examples of other techniques for detecting point mutations include, but are not limited to, selective oligonucleotide hybridization, selective amplification, or selective

primer extension. For example, oligonucleotide primers may be prepared in which the known mutation is placed centrally and then hybridized to target DNA under conditions which permit hybridization only if a perfect match is found (Saiki *et al.* (1986) *Nature, 324*:163); Saiki *et al.* (1989) *Proc. Natl. Acad. Sci. USA, 86*:6320). Such allele-specific oligonucleotides are hybridized to PCR amplified target DNA.

Alternatively, allele specific amplification technology that depends on selective PCR amplification may be used in conjunction with the instant invention. Oligonucleotides used as primers for specific amplification may carry the mutation of interest in the center of the molecule (so that amplification depends on differential hybridization) (Gibbs *et al.* (1989) *Nucleic Acids Res., 17*:2437-2448) or at the extreme 3' end of one primer where, under appropriate conditions, mismatch can prevent, or reduce polymerase extension (Prossner (1993) *Tibtech, 11*:238). In addition it may be desirable to introduce a novel restriction site in the region of the mutation to create cleavage-based detection (Gasparini *et al.* (1992) *Mol. Cell Probes, 6*:1). It is anticipated that in certain embodiments amplification may also be performed using Taq ligase for amplification (Barany (1991) *Proc. Natl. Acad. Sci. USA, 88*:189). In such cases, ligation will occur only if there is a perfect match at the 3' end of the 5' sequence making it possible to detect the presence of a known mutation at a specific site by looking for the presence or absence of amplification. Single base extension (SBE) and SBE fluorescence resonance energy transfer (SBE-FRET) can also be used to identify the specific nucleotide which occupies a given position in a nucleic acid molecule.

The methods described herein may be performed, for example, by utilizing pre-packaged diagnostic kits comprising at least one probe nucleic acid molecule or antibody reagent described herein, which may be conveniently used, *e.g.*, in clinical settings to diagnose patients exhibiting symptoms or family history of a disease or illness involving a gene of the present invention. Any cell type or tissue in which the gene is expressed may be utilized in the prognostic assays described herein.

The invention will now be described by the following non-limiting examples. The teachings of all references cited herein are incorporated herein by reference in their entirety.

EXEMPLIFICATION

Methods

All subjects participating in this study gave informed consent according to institutional and national standards (29). Sequence analysis was performed on 24 ARSACS patients from 17 families.

BAC/PAC DNA Preparation

Small quantities of DNA were prepared from BAC and PAC cell cultures (12.5 µl Chloramphenicol for BACS, 30 µg/ml Kanamycin for PACs) using a modified alkaline lysis procedure according to a published protocol (30). Larger quantities of DNA for the construction of libraries and direct sequencing were prepared using Qiagen (Qiagen, Valencia, CA) or Nucleobond columns (The Nest Group, Southboro, MA) according to the manufacturers' protocols.

M13 Library Construction and Preparation of M13 Single-Stranded DNA

BAC and PAC DNA was sheared in a sonicator to an average size of 2 kb and the ends were made blunt with Mung Bean Nuclease (New England Biolabs, Beverly, MA). The fragments were gel-purified, and subcloned into an M13mp18 *Sma* I-cut dephosphorylated cloning vector (Amersham, Uppsala, Sweden). Ligation reactions were transformed into XL2-Blue competent cells (Stratagene, LaJolla, CA). Phage plaques of M13 subclones from the BACs and PACs were grown overnight in 0.5 ml of 2x YT media with 10 µl of log phase TG-1 cells. Single-stranded M13 DNA for sequencing was purified from 100 µl of the culture supernatant with magnetic beads (PerSeptive Diagnostics, Cambridge, MA) according to the manufacturer's instructions.

Sequencing

Fluorescent sequencing of PCR products and M13 single-stranded DNA was accomplished using the Dye Primer Cycle Sequencing Ready Reaction kit (Perkin Elmer, Foster CIty, CA). Sequencing reactions contained approximately 400 ng of template in 1.5 µl and 3 µl of assay mixture for each primer. The thermal cycling parameters for the sequencing reactions were: 96°C for 10 seconds, 55°C for 5 seconds, and 70°C for 1 minute (15 cycles) followed by 96°C for 10 seconds, and 70°C for 1 minute (15 cycles). Reaction products for each primer were combined and purified with an ethanol precipitation. Sequence samples were prepared, loaded, and run on ABI 377 sequencers according to the manufacturer's instructions (Perkin Elmer). The sequences were assembled into contigs and analyzed with the STADEN software package (version 1997.1) (31, 32) and Auto Assembler (version 2.0) (Perkin Elmer). Direct sequencing of BACs was accomplished with Dye Terminator chemistry according to a previously published protocol (33). The sequence of the entire mouse and human ORFs was verified by either sequencing unambiguously on both strands or by sequencing a single strand with both the Dye Primer and the Dye Terminator reaction systems. All sequences were compared with GenBank databases and dbEST using the Search Launcher Batch Client software for Macintosh from Baylor College of Medicine (34) with Repeat/Masker pre-screening.

Computational Analyses

Web-based sequence analysis included (using default parameters):

BLAST:http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-nesblast?Jform=1;

FASTA:http://www.ebi.acc.uk/searches/fasta.html;

PSORT:http://psort.nibb.acc.jp:8800;

EXPASY Proteomics tools:http://www.expasy.ch/tools/;

BCM Search Launcher:http://www.hgsc.bcm.tmc.edu/SearchLauncher/;

mac-search-launcher:ftp://dot.bcm.tmc.edu/pub/software/search-launcher/;

COILS (35) web server:http://www.ch.embnet.org/software/COILS_form.html.

Mutation Analysis

50 ng of genomic DNA, extracted from peripheral blood leukocytes, was amplified using the primers in Figure 7. Primer pairs were designed using the web-based version of the Primer 3.0 program and PCR reactions were individually optimized. The resulting products were purified using magnetic beads (PerSeptive Diagnostics) according to the manufacturer's instructions and sequenced as above.

RNA Preparation and Northern Blot Analysis

Total RNA was extracted using the guanidinium/CsCl method from skin fibroblast cell lines from ARSACS patients and a control individual; the cell lines were grown in Eagle modified MEM (CellGro, Herndon, VA) with 10% FBS (Canadian Life Technology, Burlington, Ontario). 10 µg of RNA was electrophoresed in a 1% agarose gel and then transferred to a nylon membrane (Magna Charge, MSI, Westboro, MA) by capillary transfer with 20x SSC buffer. Pre-transfer alkaline hydrolysis of the gel was performed with 0.05M NaOH. The ^{32}P-labeled *spastin* probe was generated by random priming with the Rediprime II system (Amersham) using the 1.8 kb insert from an IMAGE cDNA clone (279258) purified after separation on low melting point agarose (Life Technologies, Rockville, MD). Hybridization for both the fibroblast blot and the multiple tissue northern blot (MTN, Human I #7760-1, Clontech) was done in ExpressHyb buffer (Clontech, Palo Alto, CA) followed by washing according to manufacturer's instructions. The size standard for both northern blots was a 0.24-9.5 kb RNA ladder (Life Technologies).

RT-PCR

500 ng of total RNA from skin fibroblasts of ARSACS patients and controls, as well as a commercial preparation of total RNA from cerebellum (Clontech), were amplified using sense and antisense primers (Figure 7) and the Superscript one step kit (Life Technologies). In all cases a parallel control reaction was set up in the absence of RT. The resulting products were purified and sequenced as above.

In situ hybridizations

Oligonucleotides complementary to nucleotides 11,009-11,055 of the human *spastin* gene (probe NIB226-1) and a sense 45-mer for the same region were synthesized and purified (MedProbe, Oslo, Norway). To exclude the possibility of any cross-hybridization to other human mRNAs, homology searches were carried out. A database search revealed no significant homologies, except for the intended targets. The oligonucleotides were subsequently labeled with a ^{35}S-labeled dATP (NEG 034H, NEN DuPont, Boston, MA) at the 3' end using terminal dideoxy nucleotidyl transferase to a specific activity of 2 x 10^9 cpm/μg and purified on a Nensorb 20 column.

The tissue was cut to 14 μm thickness in a cryostat, thawed onto Fisher probe on (+) slides (Fisher Biotech, Springfield, NJ), and processed for *in situ* hybridization according to Schalling *et al* (36). In brief, sections were incubated at 42°C for 15-18 hours with 10^6 cpm of labeled probe per 100 μl of a solution containing 50% formamide, 4x SSC, 1x Denhardt's solution, 1% sarcosyl, 0.02 M sodium phosphate ($NaPO_4$, pH 7.0) and 10% dextran sulfate mixed with 500 μg/ml sonicated salmon sperm DNA and 200 mM dithiothreitol. Sections were rinsed in 1x SSC at 55°C for one hour, dried and exposed to x-ray film (Amersham Hyperfilm β-max) for 14-21 days.

Mouse BAC Clone and Radiation Hybrid (RH) Panel Analysis

The clone containing the mouse genomic sequence (418_B_11) is from a 129 SV mouse BAC library, CitbCJ7B cloned in the vector pBeloBAC11 (Research Genetics, Huntsville, AL). The RH mapping of mouse *spastin* was performed using the T31 mouse-hamster hybrid mapping panel (11). The initial attempts with several mouse *spastin* primers failed due to the amplification of a hamster PCR product of similar size to the mouse product. A hamster PCR product was sequenced, which revealed minor sequence differences with mouse *spastin*. The successful mouse *spastin* primers were MARS-3F ((TCATTCATATGTCCCAGGGACATGT; SEQ ID NO: 72) and MARS-3R (CTACTAGAACTGCATGTGCCGC; SEQ ID NO: 73). The RH vector obtained from

testing the T31 panel was compared to the reference map generated at MIT (12) using the "placement" function of RHMAPPER.

Computation of the P_{excess} Statistic

Seven-marker haplotypes for 55 ARSACS and 58 normal chromosomes were obtained from 68 obligate carrier parents by not counting copies that were considered to be identical by descent within a pedigree (5). Marker haplotypes were constructed using the SIMWALK2 program (37). The simple linkage disequilibrium mapping measure $P_{excess}=(p_{affected} - p_{normal})(1-p_{normal})$, 23, 38, 39, 40) was calculated from the frequencies of haplotypes.

While this invention has been particularly shown and described with references to preferred embodiments thereof, it will be understood by those skilled in the art that various changes in form and details may be made therein without departing from the scope of the invention encompassed by the appended claims.

References

1. Bouchard J-P. Recessive spastic ataxia of Charlevoix-Saguenay. *In*: "Handbook of Clinical Neurology 16: hereditary neuropathies and spinocerebellar degenerations", (J.M.B.V. de Jong, Ed.) pp. 451-459, Elsevier Science Publishers, Amsterdam. (1991).
2. Bouchard, J.P., *et al*. Autosomal recessive spastic ataxia of Charlevoix-Saguenay. *Neuromuscular Disorders* **8**, 474-479 (1998).
3. De Braekeleer, M., *et al*. Genetic epidemiology of autosomal recessive spastic ataxia of Charlevoix-Saguenay in northeastern Quebec. *Genetic Epidemiology* **10**, 17-25 (1993).
4. Charbonneau H. & Robert N. The French origins of the Canadian population 1608-1759. In: Harris RC (ed) Historical atlas of Canada Volume I: from the beginning to 1800, University of Toronto Press, Toronto, plate 45 (1987).
5. Richter, A., *et al*. Location score and haplotype analyses of the locus for autosomal recessive spastic ataxia of Charlevoix-Saguenay in chromosome region 13q11. *Am. J. Hum. Genet* **64**, 768-775 (1999).
6. Engert, J.C., *et al*. High Resolution Physical and Transcript Map of the Autosomal Recessive Spastic Ataxia of Charlevoix-Saguenay (ARSACS) Candidate Region in Chromosome 13q11. *Submitted* (1999).
7. Fink, AL. Chaperone-mediated protein folding. *Physiological Reviews*. 79, 425-49 (1999).
8. Buchner J. Hsp90 & Co. - a holding for folding. *Trends in Biochemical Sciences*. **24**, 136-41, (1999).
9. Gupta, R. S. Phylogenetic analysis of the 90 kD heat shock family of protein sequences and an examination of the relationship among animals, plants, and fungi species. *Molecular Biology & Evolution*; 12, 1063-1073 (1995).
10. Nakai, K. & Kanehisa, M. A knowledge base for predicting protein localization sites in eukaryotic cells, *Genomics* **14**, 897-911 (1992).
11. McCarthy, L.C., *et al*. A First-Generation Whole-Genome Radiation Hybrid Map Spanning the Mouse Genome. *Genome Research* 7, 1153-1161 (1997).
12. W.J. Van Etten, *et al*. Radiation hybrid map of the mouse genome. *Nature Genet.* 22, 384-387 (1999).
13. Brown CJ, *et al*. The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*. 71, 527-42, (1992).

14. Porchet, N, Aubert, JP, & Laine, A. MUC5B, the 10.7-kb large central exon encodes various alternate subdomains resulting in a super-repeat. *Journal of Biological Chemistry* 272, 3168-78 (1997).

15. Gentles AJ. & Karlin S. Why are human G-protein-coupled receptors predominantly intronless? *Trends in Genetics*. 15, 47-49 (1999).

16. Palmer JD. & Logsdon JM Jr The recent origins of introns. *Current Opinion in Genetics & Development*. 1, 470-7, (1991).

17. Edward M. Marcotte *et al.* Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. *Science* 285, 751-753 (1999).

18. Prodromou C. et al. Identification and structural characterization of the ATP/ADP-binding site in the Hsp90 molecular chaperone. *Cell* 90, 65-75 (1997).

19. Kimura Y. Yahara I. & Lindquist S. Role of the protein chaperone YDJ1 in establishing Hsp90-mediated signal transduction pathways. *Science* 268, 1362-1365, (1995).

20. Dittmar KD. Banach M. Galigniana MD. & Pratt WB. The role of DnaJ-like proteins in glucocorticoid receptor.hsp90 heterocomplex assembly by the reconstituted hsp90.p60.hsp70 foldosome complex. *Journal of Biological Chemistry* 273, 7358-66, (1998).

21. Dickie, M.M. Tumbler, tb, *Mouse News Lett*;32, 45 (1965).

22. L Kruglyak. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.*22, 139-144 (1999).

23. Hästbacka J., *et al.*. Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genet.* 2, 204-211 (1992).

24. Thompson EA. & Neel JV. Allelic disequilibrium and allele frequency distribution as a function of social and demographic history. *American Journal of Human Genetics*. 60, 197-204, (1997).

25. Graham J. & Thompson EA. Disequilibrium likelihoods for fine-scale mapping of a rare allele. *American Journal of Human Genetics*. 63, 1517-30, (1998).

26. Boehnke M. Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. *American Journal of Human Genetics*. 55, 379-390 (1994).

27. Graham J. (thesis) Disequilibrium fine-mapping a rare allele via coalesent models of gene ancestry Ph.D., Univ. of Washington, Seattle (1998).

28. De Braekeleer, M. Geographic distribution of 18 autosomal recessive disorders in the French Canadian population of Saguenay-Lac-St-Jean, Quebec. *Annals of Human Biology* 22, 111-122 (1995).

29. Knoppers BM, & Laberge C DNA sampling and informed consent. *Can Med Assoc J* **144**, 128-129 (1991).

30. Birren, B.W., Mancino, V., & Shizuya, H. Bacterial Artificial Chromosomes. *In* "Genome Analysis: A Laboratory Manual" Volume 3 (Birren, B., Green, E.D., Klaphoz, S., Myers, R. M., Riethman, H., and Roskams, J. Eds.), pp. 241-295, Cold Spring Harbor Laboratory Press, Plainview NY (1999).

31. Bonfield JK. Smith Kf. & Staden, R. A new DNA sequence assembly program. *Nucleic Acids Research*. **23**, 4992-4999 (1995).

32. Bonfield JK. & Staden, R. Experiment files and their application during large-scale sequencing projects. *DNA Sequence* **6**, 109-117, (1996).

33. Boysen C., Simon MI, & Hood L. Fluorescence-based sequencing directly from bacterial and P1-derived artificial chromosomes. *Biotechniques* **23**, 978-82 (1997).

34. Smith RF, Wiese BA, Wojzynski MK, Davison DB, & Worley KC. BCM Search Launcher--An Integrated Interface to Molecular Biology Data Base Search and Analysis Services Available on the World Wide Web. *Genome Res* **6**, 454-62 (1996).

35. Lupas, A., M. Van Dyke, & J. Stock, Predicting Coiled Coils from Protein Sequences. *Science* **252**, 1162-1164 (1991).

36. Schalling M. *et al* Neuropeptide Y and catecholamine synthesizing enzymes and their mRNAs in rat sympathetic neurons and adrenal glands: Studies on expression, synthesis and axonal transport after pharmacological and experimental manipulations using hybridization techiques and radioimmunoassay. *Neuroscience* **41**, 753-766 (1991).

37. Weeks DE, Sobel E, OfConnell JR, & Lange K. Computer programs for multilocus haplotyping of general pedigrees. *Am J Hum Genet* **56**, 1506-1507 (1995).

38. Devlin B. & Risch N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311-322 (1995).

39. de la Chapelle, A. & Wright, F. A.D. Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. *Proc. Natl. Acad. Sci., USA* **95**, 12416-23 (1998).

40. Austerlitz F. & Heyer E. Impact of demographic distribution and population growth rate on haplotypic diversity linked to a disease gene and their consequences for the estimation of recombination rate: example of a French Canadian population. *Genetic Epidemiology*. **16**, 2-14, (1999).

41. McNally, E.M., *et al*.. Mild and severe muscular dystrophy caused by a single gamma-sarcoglycan mutation. *Am. J. Hum. Genet.* **59**, 1040-1047 (1996).

42. Nagase,T., *et al*... Prediction of the coding sequences of unidentified human genes. XI. The complete sequences of 100 new cDNA clones from brain which code for large proteins in vitro. *DNA Res.* 5, 277-286 (1998).

43. Thompson, J.D., Higgins, D.G. & Gibson, T.J.. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22, 4673-4680 (1994).